Improving the Resolution of Retinal OCT with Deep Learning

Ying Xu[1], Bryan M. Williams[2], Baidaa Al-Bander[1], Zheping Yan[3], Yao-chun Shen[1], and Yalin Zheng[2]

[1] Department of Electrical Engineering and Electronics, University of Liverpool, L69 3GJ, UK
[2] Department of Eye and Vision Science, University of Liverpool, L7 8TX, UK, yzheng@liverpool.ac.uk,
WWW home page: http://www.liv-cria.co.uk
[3] College of Automation, Harbin Engineering University, Harbin, Heilongjiang 150001, China

Abstract. In medical imaging, high-resolution can be crucial for identifying pathologies and subtle changes in tissue structure. However, in many scenarios, achieving high image resolution can be limited by physics or available technology. In this paper, we aim to develop an automatic and fast approach to increasing the resolution of Optical Coherence Tomography (OCT) images using the data available, without any additional information or repeated scans. We adapt a fully connected deep learning network for the super-resolution task, allowing multi-scale similarity to be considered, and create a training and testing set of more than 40,000 sample patches from retinal OCT data. Testing our model, we achieve an impressive root mean squared error of 5.847 and peak signal-to-noise ratio (PSNR) of 33.28dB averaged over 8282 samples. This represents a mean improvement in PSNR of 3.2dB over nearest neighbour and 1.4dB over bilinear interpolation. The results achieved so far improve over commonly used fast techniques for increasing resolution and are very encouraging for further development towards fast OCT super-resolution. The ability to increase quickly the resolution of OCT as well as other medical images has the potential to impact significantly on medical imaging at point of care, allowing significant small details to be revealed efficiently and accurately for inspection by clinicians and graders and facilitating earlier and more accurate diagnosis of disease.

Keywords: Super Resolution, Retina OCT, Fully Convolutional Networks

1 Introduction

Optical coherence tomography (OCT) has emerged as one of the most common scans performed in ophthalmology. With a high spatial resolution of better than $10\mu m$ [9], OCT is capable of achieving in-vivo tomographic scans of tissue in fine detail. With this high-resolution capability, OCT allows for early diagnosis and

tracking of the progression of ocular diseases via visualisation of the intraretinal and intracorneal architectural morphology [4]. The most prominant technique to achieve this resolution is the combination of a high-speed spectrometer with short confocal parameter to alleviate the focus secession [2,8]. Despite its potential to reveal vital pathological clues, the current achievable resolution of OCT is not sufficient to visualise small structures such as Descemet's membrane and Bowman's layer in the cornea.

With SR algorithms, we aim to break the limit on OCT resolution imposed by the diffraction of light in the optical path [4,9]. There have been many methods reported in the literature aiming to achieve super-resolution. Methods for OCT aim to improve resolution by using texture priors and interpolation [5]. While these may achieve reasonable and fast results, they can have a smoothing effect on structure edges, regions and their contents. Lyndsey et al. [10] proposed a learning-based method using the idea that the local geometry of feature spaces in low-resolution (LR) patches are analogous to the corresponding manifolds in high-resolution (HR) patches. Radu et al. [13] proposed a neighbour embedding method using a sparse coding method to predict the corresponding high-resolution embedding matrix. This type of method is capable of recovering sharp images with fine details, but at a computational and time cost.

Since super-resolution is a nonlinear process and neural networks specialise in fitting nonlinear mappings, some recent work has considered the application of neural networks to super-resolution. Kwang et al. [6] used neighbour embedding to exploit kernel ridge regression and gradient descent to ascertain localisation accuracy. Dong et al. [3] employed a convolutional network to achieving super-resolution from the LR counterparts of given HR images. This model reconstructs SR output only in the last layer, suggesting that detailed information could be lost for larger magnification tasks. A recent SR method based on the generative adversarial network (GAN) [7] further developed some countermeasures to recovering missing texture information.

Unlike classical CNNs, we explore an encoder-decoder type fully convolutional network (FCN) accepting inputs of arbitrary size. Such FCN-based networks have been widely applied in segmentation tasks, however they have not been well studied in the case of super-resolution. Differing from traditional CNN methods [3], the fully convolutional layer is substituted with several upsampling layers which can be regarded as a gradual multi-step magnification. Compared with the gradual up-scaling generative model [7], FCNs also reveal a possible way to generate feature maps while preserving the spatial information from the contracting path.

In this work, we propose a novel approach to super-resolution. Building upon U-net [11], we present a novel architecture called SRUnet, with further embedded gradual increases in resolution. With this network, we aim to capture both the task-relevant visual details and scale-invariant image structures of OCT retinal images. We make the following contributions: (i) We develop a new neural network SRUnet, based on UNet, for super-resolution and extend this to varying magnification strengths using successive deconvolution stages to

achieve this reliably. (ii) We develop a series of large datasets from retinal OCT images, allowing for the model to be trained, validated and tested. (iii) We test our model at various resolutions and for various magnification strengths.

2 Methods

Our aim is to achieve super-resolution by recovering high-resolution patches from corresponding low-resolution input. In order to obtain suitable datasets for testing, we first produce a series of overlapping patches P^{HR} from high-resolution OCT scans I^{HR} such that every image pixel appears in at least one patch. We then reduce the resolution of the patches by downsampling by nearest-neighbour interpolation to achieve the set of corresponding low-resolution patches P^{LR}. This simulates the idea of missing photons in an imaging system and allows us to have ground truth datasets for training and comparison. Using these, the neural network can attempt to learn the link between low-resolution and high-resolution OCT image patches. To achieve several datasets and to allow for testing with various resolutions and magnifications, we select patches of size $W \times H$, and we downsample by a factor of r to obtain sizes $H/r \times W/r$. The patches are then fed into our neural network which attempts to learn the correspondence between HR and LR training pairs. Once trained, the network is tested on the set of LR testing patches and corresponding SR patches are produced. Finally, we reconstruct the recovered SR patches into whole B-scan images using the mean value on the overlapping regions. See the processing portion shown in Fig. 1.

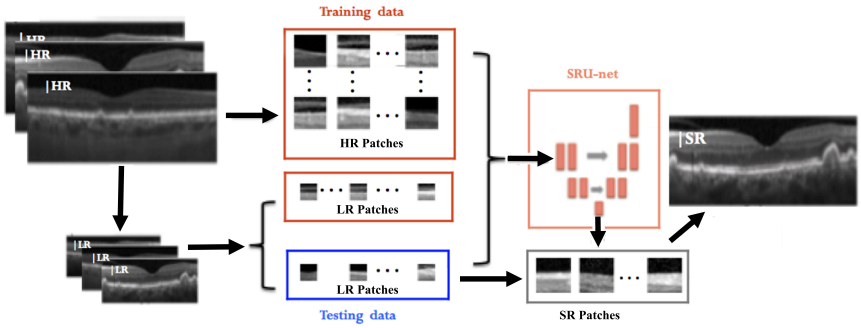

Fig. 1. Flow chart of our overall approach. We train the network with corresponding low- and high-resolution image patches. In order to test, we break the low-resolution OCT images into overlapping patches, obtain corresponding SR patches from our trained network and reconstruct the whole SR OCT images.

2.1 Network Architecture

We use Unet [11] as the main basis of our architecture. The basic schematic of our SRUnet network is shown in Fig.2. The proposed pixel-wise, encoder-decoder

network, which introduces several deconvolution layers, creating a gradually deeper network to estimate HR detail. SRUnet is capable of taking various input sizes and producing high-resolution results. At the training stage, patches from both I^{HR} and I^{LR} are used as input. Let R_n denote a Convolution-LeakyReLU-BatchNorm layer [14] with n feature maps. In the encoder, the total contracting path is $R_{32} - R_{64} - R_{128} - R_{256} - R_{512} - R_{512} - R_{512}$. After quadruplication of each max-pooling layer (stride=2), the highest spatial-frequency counterparts of I^{HR} are captured. The decoder uses similar functions without batch normalisation and replaces the activation function with ReLU. Several upsampling layers (stride=0.5) automatically discover high-dimensional manifolds and embed them to concatenating low-dimensional feature maps with zero padding. The sequence of extra sub-pixel deconvolutional layers, as mentioned by [12], is the key aspect which facilitates determination of the SR patch. After the last deconvolutional layer, a fully connected layer is applied to map a one-dimensional output, followed by a Sigmoid function. The encoder employs dropout from the third layer with a rate of 0.3 and the slope of LeakyReLU is 0.2. All convolutions are 1×1 with a stride of 2.

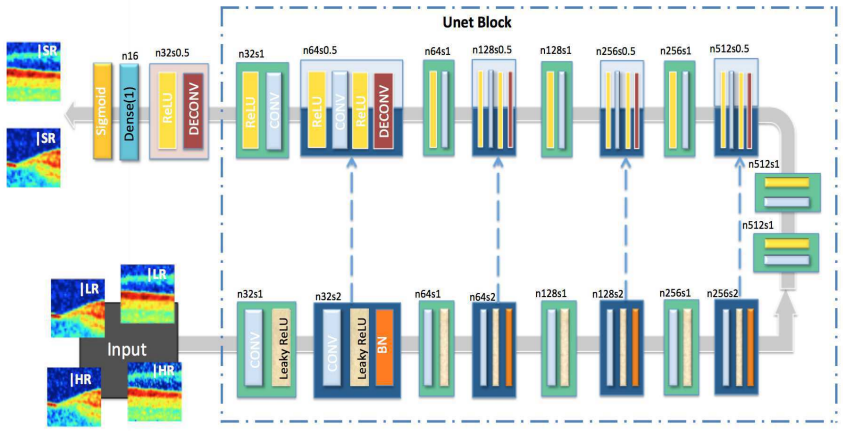

Fig. 2. Proposed SRUnet architecture with corresponding kernel size (k), number of feature maps (n) and stride (s) indicated for each convolutional layer

2.2 Upsampling Layer Checking

In FCNs, the deconvolution component of the network is generally referred to as upsampling. It typically relies on bilinear interpolation to increase the size of the feature maps of the previous layer. The output is embedded to have the same size as the ground truth so that pixel-wise error measures may be calculated. The local texture features can be characterised by the gradually tuned weights. Particularly, we do not apply nonlinearity to the outputs of the

last deconvolution layer, since this would reduce the resolution of the recovered HR feature maps.

2.3 Cost Function

To learn a nonlinear mapping from the LR image patches to corresponding high-dimensional patches, the vectors consist of a dedicated set of feature maps. Let P^{LR} be the input low-resolution image patch set and let their corresponding high resolution patches be denoted by P^{HR}, the learning mapping function S along with parameters $\theta = \{W_1 : N; B_1 : N\}$, where W_1 and B_1 denote the Nth layer's weights and biases respectively. We use cross entropy as a cost function in our SRUnet, given by:

$$\mathrm{H(S)} = \mathbb{E}_{I^{HR} \sim \boldsymbol{p}_{\mathrm{train}}(I^{HR})} \left[-\sum_{i}^{n} I_i^{HR} \log \left(S_{\theta_i} \left(I_i^{LR} \right) \right) \right] \tag{1}$$

where n denotes the total number of pixels and i denotes the pixel index. This is used to allow the reconstruction of images closely resembling the ground-truth image data. We also calculate the pixel-wise Mean Square Error (MSE) so that the performance of the network can be evaluated in real time during training.

3 Experimental Results

In this section, we test the performance of our network to improve the resolution of LR images by 2x with the fixed SRUnet architecture. Further to this, we investigate the effect of the adjusting the depth of deconvolutional layers, allowing for greater magnification than the basic SRUnet. Specifically, Fig. 2 shows our basic SRUnet with a single additional deconvolution block for a 2x increase in resolution. To increase the magnification abilities, we introduce additional deconvolutional steps, keeping the small stride step invariant.

Dataset: We test our approach using a set of 320 B-scan OCT images from a Diabetic Retinopathy study [1]. These images were acquired from 34 patients, 30 of which had both eyes scanned, while four had only one eye scanned. For each eye, five images were captured at different intervals. Overlapping patches were taken from the 1025×496 B-scan OCT images with sizes $256x256$, $128x128$ and $64x64$. These were then downsampled using nearest neighbour interpolation to $128x128$, $64x64$ and $32x32$, allowing us to create 6 datasets: three with 2x magnification, two with 4x magnification and one with 8x magnification. We partitioned the datasets at random into training (80%) and testing (20%) sets; of each training dataset, 20% was reserved for validation.

Implement Details: The implemented network was built on the Keras framework with a Tensorflow backend, using CUDA (NVIDIA Corp). All initial weights were randomly drawn from a Gaussian distribution of mean 0 and standard deviation 0.01. Each model was trained until no further improvement of the cost

function was achieved and then predictions were generated for the testing set. The initial learning rate was kept at be 10^{-4}. All tests were run on a single NVidia Titan Xp GPU; the inference took less than half a second per image.

Experimental setup: For these experiments, we consider varying patch sizes n and depths of deconvolutional layers. We evaluate performance using Peak Signal to Noise Ratio (PSNR). The images I^{HR} were cropped to squares of $2^n \cdot 64 \times 2^n \cdot 64$ pixels, where n is the scale expansion index ($n = 0, 1, 2$). The corresponding low-resolution patch P^{LR} is related by the down-sampling factor r. For the SRU_γ network, the number of deconvolutional layers needed is $\gamma = \log_{\frac{1}{s}} r$, where stride step (s) is fixed at 0.5 in our upsampling operation. Examples of our predicted SR patches and corresponding ground truth patches are shown in Fig.3.

Using 8282 randomly selected samples from our dataset of 32×32 images, we achieved average root mean squared error of 5.847 and peak signal-to-noise ratio (PSNR) of 33.28dB for 2x magnification. This corresponded to a mean improvement in PSNR of 3.2dB over nearest neighbour and 1.4dB over bilinear interpolation. We also tested the network sensitivity to different HR patch sizes and magnifications using the full datasets. With an upsampling factor of 2, the 32×32 images achieved a mean PSNR of 33.45, the 64×64 set achieved a higher PSNR of 33.79 and the 128×128 set achieved 33.94. While, as expected, patch size affected the results slightly there was no significant change. The training time for larger images increases considerably but the very fast testing time is not increased significantly. Increasing to 4x magnification reduces the quality of the results slightly with PSNRs of 30.31 and 30.85 for the 32×32 and 64×64 sets respectively and it is reduced further for the very challenging task of 8x magnification, achieving 27.96 for the 32×32 images.

Number of layers: Larger magnification tasks can be achieved by involving more multi-step upsampling layers. We try deeper structures by adding at most another three non-linear mapping layers to propagate local information to a broader region. We compare the achieved SR images in Fig.3 for $2\times$, $4\times$ and $8\times$ magnification. We can observe all predicted image recovered similar structure as the ground truth within a strong shape variation. In addition, both $2\times$ or $4\times$ provide easier access to learn sharp edges with stable local structure. However, networks involving more layers converge slower under the same learning rate. The performance of SRUnet is also limited by the quality of a coarse LR image, high dimensional textural features can be lost despite deeper learning.

4 Discussion

We designed a fully connected neural network SRUnet based on Unet to achieve image super-resolution. We validated and tested three variants of the network for improving OCT resolution from OCT B-scan images. The SRUnet models yield an accuracy of 30+ dB on the test set, which outperforms nearest neighbour interpolation (NNI) and bicubic interpolation (BI). The comparison with

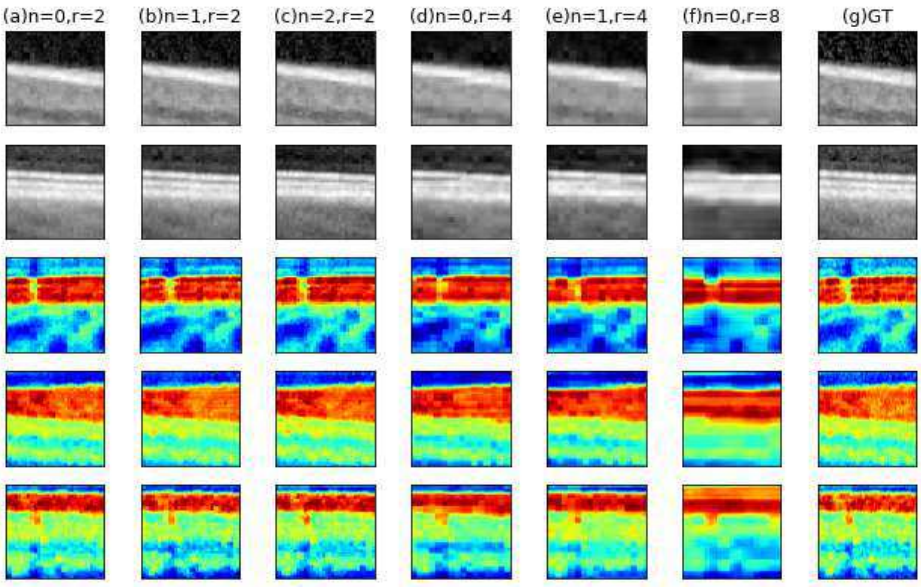

Fig. 3. Examples of our test results. Columns (a)-(c) show examples of 2x SR, columns (d)-(e) show 4x SR, column (f) shows 8x SR and (g) shows the ground truth HR patch which we aim to approximate

ground truth demonstrates the network's accuracy. It is also demonstrates that noise was suppressed while the sharp image contours were preserved; however, we do not have ground truth information to evaluate this and denoising was not our primary goal. In contrast to other algorithms based on iteration, our algorithm is computationally fast. However, the loss function used appears to give overly smooth results. There may be concern that the neural network is learning patient specific information but this is a common concern with machine learning algorithms and can be addressed by building larger datasets. While we achieved good results for 2x and 4x magnification, we started to see limitations with larger 8x magnification. In future work, we aim to improve on this, extending it to fundus images and explore the performance in images with known pathologies.

5 Conclusion

In this paper, we have proposed SRUnet, a novel FCN framework to improve the resolution of OCT images, increasing both PSNR and visual quality. We obtained super-resolution images with little computational time cost, meaning that our method could be implemented in real time.

References

1. Boonarpha, N.: Choroidal structure and function in chronic retinal diseases. Doctoral dissertation, University of Liverpool (2016)
2. Boppart, S.A., Herrmann, J., Pitris, C., Stamper, D.L., Brezinski, M.E., Fujimoto, J.G.: High-resolution optical coherence tomography-guided laser ablation of surgical tissue. J. Surg. Res. 82(2), 275–284 (1999)
3. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE T. Pattern Anal. 38(2), 295–307 (2016)
4. Fujimoto, J., Boppart, S.A., Tearney, G., Bouma, B., Pitris, C., Brezinski, M.: High resolution in vivo intra-arterial imaging with optical coherence tomography. Heart 82(2), 128–133 (1999)
5. Gargesha, M., Jenkins, M.W., Wilson, D.L., Rollins, A.M.: High temporal resolution oct using image-based retrospective gating. Opt. Express 17(13), 10786–10799 (2009)
6. Kim, K.I., Kwon, Y.: Single-image super-resolution using sparse regression and natural image prior. IEEE T. Pattern Anal. 32(6), 1127–1133 (2010)
7. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. arXiv preprint (2016)
8. Miller, D., Kocaoglu, O., Wang, Q., Lee, S.: Adaptive optics and the eye (super resolution oct). Eye 25(3), 321 (2011)
9. Nassif, N., Cense, B., Park, B., Pierce, M., Yun, S., Bouma, B., Tearney, G., Chen, T., De Boer, J.: In vivo high-resolution video-rate spectral-domain optical coherence tomography of the human retina and optic nerve. Opt. Express 12(3), 367–376 (2004)
10. Pickup, L.C., Roberts, S.J., Zisserman, A.: A sampled texture prior for image super-resolution. In: Advances in neural information processing systems. pp. 1587–1594 (2004)
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
12. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1874–1883 (2016)
13. Timofte, R., De, V., Van Gool, L.: Anchored neighborhood regression for fast example-based super-resolution. In: Computer Vision (ICCV), 2013 IEEE International Conference on. pp. 1920–1927. IEEE (2013)
14. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)